On the Dangers of Modelling through Continuous Distributions

Fernández, C.; Steel, M.F.J.

Publication date:
1997

ON THE DANGERS OF MODELLING THROUGH CONTINUOUS DISTRIBUTIONS: A BAYESIAN PERSPECTIVE

By Carmen Fernández[1] and Mark F.J. Steel

CentER for Economic Research and Department of Econometrics
Tilburg University, 5000 LE Tilburg, The Netherlands

First Version OCTOBER 1996; Present Version JANUARY 1997

SUMMARY

We point out that Bayesian inference on the basis of a given sample is not always possible with continuous sampling models, even under a proper prior. The reason for this paradoxical situation is explained, and its empirical relevance is linked to coarse gathering of data, such as rounding. A solution, inspired by the way observations are recorded, is proposed. Use of a Gibbs sampler makes the solution practically feasible. The case of independent sampling from (possibly skewed) scale mixtures of Normals is analysed in detail for a location-scale model with a commonly used noninformative prior. For Student-t sampling with unrestricted degrees of freedom the "usual" inference, based on point observations, is shown to be precluded whenever the sample contains repeated observations. We show that Bayesian inference based on set observations, however, is possible and illustrate this by an application to a skewed data set of stock returns.

Keywords: COARSE DATA; POSTERIOR EXISTENCE; LOCATION-SCALE MODEL; ROUNDING; SCALE MIXTURES OF NORMALS; SKEWNESS; STUDENT-t

1. INTRODUCTION

The analysis of rounded or grouped data through the use of continuous sampling models has permeated the statistical literature for almost a century. Ever since Sheppard (1898) proposed his correction, the issue of how inference is affected by the use of "coarse" data has been an integral part of statistics. Sheppard's correction was given a likelihood justification for small rounding error in Fisher (1922) and Lindley (1950) under univariate Normal sampling. Dempster and Rubin (1983) provide a maximum-likelihood foundation for this correction in a wider context, extending to multivariate Normality and large samples from "regular" models. An overview of the literature is provided in Heitjan

[1] *Address for correspondence:* Department of Econometrics, Tilburg University, P.O.Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: carmen@kub.nl

(1989). In the more complicated case where the actual coarsening mechanism (leading to *e.g.* heaping or censoring) is itself random, Heitjan and Rubin (1991) analyse when the random nature of the coarsening can be ignored: they characterize this situation as "coarsening at random" (CAR). An area of application where the effects of rounding have received considerable attention is finance; typically, prices of securities are measured in discrete units. Ball (1988) investigates Sheppard's correction in this context and proposes a maximum-likelihood estimator for the variance. Hausman, Lo and MacKinlay (1992) use a more general ordered probit model to capture this discrete feature of stock prices.

In this paper, we shall consider situations where CAR applies, and, in the interest of readability, most of the paper will be centered around the simple case of rounding. Extending the discussion to any other type of CAR is trivial, however. The focus of the literature in this area has, to our knowledge, been the *quantitative* effect of coarsening on inference. This paper, on the other hand, examines the *qualitative* effect of coarsening on Bayesian inference. In particular, we point out situations where inference would be possible on the basis of data in accordance with the continuous sampling model, but not on the basis of recorded point observations. Thus, the "observed" sample leads to an improper posterior and inference is out of the question, even with a proper prior!

The source of such pathological behaviour lies in the fact that, in the real world, observations are recorded as numbers, whereas continuous sampling models always give probability zero to any such number. Thus, point observations are formally in conflict with the sampling assumptions. Furthermore, the conditional distribution of the parameters given the observables (*i.e.* the posterior distribution) can fail to be defined for a set of measure zero in the observables. At first sight, this does not seem troublesome, since any problem can only occur on a set of samples which has probability zero of being generated by the sampling model. However, since any recorded sample has probability zero of occurrence, we can never be sure that the sample under consideration is not an "offending" one. What makes this problem of practical relevance is that rounding can give a nonnegligible probability to an "offending" sample actually being recorded. If, *e.g.* given the value zero for the observable, the posterior is not well-defined, rounding the observations to a finite precision can make it quite possible to "observe" zero, even though that value has no probability attached to it by any continuous sampling process. The solution we propose here is inspired by how the data are actually observed, and identifies a point observation with the neighbourhood that would have led to this value being reported. In that case, we show that inference is always possible under a proper prior. For models with improper priors we still need to verify whether the predictive mass assigned to the particular sample of set observations we consider is finite.

Numerical methods are a double-edged sword in all this: on the one hand, problems of an improper posterior often go unnoticed whenever numerical methods are applied without any analytical verification, but on the other hand, they provide the key to the practical applicability of the solution. The use of Markov chain Monte Carlo methods, such as Gibbs sampling, renders the solution quite feasible.

As an important example, we present in detail the case of independent sampling from scale mixtures of Normals. For practical purposes, an important member of this class is the Student-t model. We also allow for extending these models to their skewed counterparts

and complement the sampling model with a commonly used improper prior. Detailed results are presented for the analyses using both point observations and set observations.

Finally, an application to stock price returns illustrates the problem and shows the empirical feasibility of the solution using set observations. In an extreme case, this analysis is seen to be far preferable to a more ad-hoc solution to the problem.

For probability density functions, we use the notation of DeGroot (1970), and all proofs are grouped in Appendix A.

2. THE FUNDAMENTAL PROBLEM

In this Section we shall explain the source of the problems one may face when conducting posterior inference with continuous sampling distributions. We assume throughout that the dominating measure is the Lebesgue measure in the corresponding space.

We thus consider a sampling distribution with probability density function (p.d.f.) $p(y|\theta)$ where $y \in \mathcal{Y} \subseteq \Re^n$ and $\theta \in \Theta \subseteq \Re^m$. We complete the Bayesian model with a σ-finite prior distribution given through a density $p(\theta)$, which could either be proper or improper. The resulting Bayesian model uniquely defines a joint σ-finite distribution on $\mathcal{Y} \times \Theta$ with density

$$p(y, \theta) = p(y|\theta)p(\theta). \tag{2.1}$$

If $p(\theta)$ is proper this joint distribution can be decomposed into the marginal (predictive) distribution of y with p.d.f.

$$p(y) = \int_{\Theta} p(y|\theta)p(\theta)d\theta, \tag{2.2}$$

and the conditional distribution of θ given y, defined through the p.d.f.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{2.3}$$

if $p(y) < \infty$ and arbitrarily otherwise. Note that since $p(y)$ in (2.2) is a p.d.f., the set of y's for which $p(y) = \infty$ has Lebesgue measure zero and thus zero probability of being observed. Therefore, the probability of observing an "offending" value of y is zero and, in principle, we need not worry about such values. Of course, under a continuous sampling distribution, *any* given value y_0 has zero probability of occurring. However, current statistical practice is to conduct inference on θ on the basis of (2.3) with y replaced by the "observed" value y_0. The fact that a value y_0 is actually recorded as "observed" data, in spite of having zero occurrence probability under the continuous sampling model, is the result of a measuring scheme, which typically assigns a particular value y_0 to the factual observation that y lies in some neighbourhood around y_0. Although this is standard practice, it can have serious implications since there is no guarantee that the reported value y_0 is not an "offending" value, corresponding to $p(y_0) = \infty$. The latter would clearly imply that $p(\theta|y_0)$ is not defined and, thus, inference on θ is precluded.

Whenever $p(y_0)$ can be computed analytically, such a problem would be detected, but the vast majority of statistical applications to complex real-life problems has to rely on numerical methods, which may well fail to indicate the problem. Curiously, common practice does not include checking whether $p(y_0) < \infty$ in models with a proper prior. Thus, there is a danger of reporting senseless inference. In this paper we intend to investigate some situations where this danger is of real practical importance.

As an illustration, we present the following simple example. More practically relevant models will be analysed in detail in Section 4.

Example 1. *A Scale Contaminated Model*
Let us consider n i.i.d. replications $(y_1, \ldots, y_n)$ from the ε-contaminated model with p.d.f.

$$p(y_i|\sigma) = (1-\varepsilon) f_N(y_i|0, \sigma^2) + \varepsilon f_N(y_i|0, c^2), \tag{2.4}$$

where $f_N(y_i|\xi, \omega^2)$ denotes the density function of a Normal distribution with mean ξ and variance ω^2 evaluated at y_i. In such a model $\varepsilon \in (0, 1/2)$ could represent the probability of y_i being an "outlying" observation, generated with variance $c^2 > 1$, whereas the "usual" observable has variance $\sigma^2 < 1$. For convenience, we shall assume both ε and c^2 fixed, but the following results carry over to the case with a proper prior on (ε, c^2). The prior assumed for σ will be a Beta(a, b) distribution with p.d.f.

$$p(\sigma) = B(a,b)^{-1} \sigma^{a-1} (1-\sigma)^{b-1} I_{(0,1)}(\sigma), \tag{2.5}$$

where $B(a, b)$ is the Beta function and I_H denotes the indicator function of the set H. Clearly, if $y_i \neq 0$ for all $i = 1, \ldots, n$, the likelihood function from (2.4) is bounded and thus leads to a finite integral under any proper prior. If, however, $r \geq 1$ observations are equal to zero, the likelihood can be shown to have upper and lower bounds both proportional to σ^{-r}. Therefore, a finite predictive density value is achieved only when $a > r$. Thus, use of the proper prior in (2.5) with $a \leq r$ (the number of zero "observations") does not allow for posterior inference.

Under an improper prior $p(\theta)$, the decomposition in (2.2) – (2.3) still applies if and only if the predictive distribution is σ-finite, *i.e.* the density $p(y)$ in (2.2) is finite except possibly for a set of y's of Lebesgue measure zero in $\Re^n$ [see Mouchart (1976) and Florens, Mouchart and Rolin (1990)]. Obviously, the danger arising from plugging in a particular value y_0 in (2.3) carries over to this case. However, since it is well-known that the use of an improper prior on θ may preclude the existence of the posterior distribution, it is then common practice to check whether $p(y)$ defined in (2.2) is finite at the "observed" value y_0. Whereas this guarantees that the expression in (2.3) with $y = y_0$ defines a p.d.f. for θ, it does not, however, imply the existence of a conditional distribution, since from $p(y_0) < \infty$ it does not follow that the predictive distribution is σ-finite. If the latter does not hold, $p(\theta|y_0)$ is properly normalized but can not be interpreted as the conditional distribution of the parameter given the observable.

To summarize the ideas explained so far, we can mention two separate issues:

Condition A. The existence of a conditional distribution of the parameter θ given the observable y.

Condition B. The fact that (2.3) defines a p.d.f. for θ given a particular "observation" y_0.

Our point is that neither Condition A nor Condition B implies the other. It may well happen that A holds (under a proper prior it always does) but still $p(y_0) = \infty$ for a certain value y_0, in which case $p(\theta|y_0)$ is not defined. Conversely, the fact that $p(y_0) < \infty$ for a given value y_0 (and thus $p(\theta|y_0)$ is a p.d.f. for θ) does not imply that a conditional distribution for θ given y exists. The ideal situation for conducting Bayesian inference is when both A and B hold simultaneously: while A provides an interpretation of the distribution of θ given y as a conditional distribution, B is clearly required if we wish to conduct inference on the basis of a point observation y_0.

These issues are thus quite distinct, yet often seem to be confused in practice. Under a proper prior, A is known to hold and checking whether $p(y_0) < \infty$ is virtually never done. This practice seems to overlook the fact that B can then still fail to hold for certain "observed" samples, thus precluding Bayesian inference on the basis of such data. The often presumed "automatic" feasibility of Bayesian inference under proper priors can, therefore, be destroyed by the use of point observations that are fundamentally incompatible with the sampling model. Furthermore, such a breakdown will often even go undetected when the analysis only relies on numerical methods. On the other hand, under an improper prior, common practice is to check whether $p(y_0) < \infty$, although the fact that A may, nevertheless, not hold is often neglected. For practical purposes, however, this seems a somewhat lesser evil, since posterior inference will at least be based on the properly normalized (2.3), which can always be given an appealing heuristic interpretation as a full description of our knowledge after "observing" y_0.

In this paper we shall be concerned with situations where A holds but B does not, as we have a point observation y_0 for which $p(y_0) = \infty$. The next Section proposes a solution that allows for Bayesian inference in this situation, through a more careful modelling of the data generating mechanism, in accordance with the way the data are actually observed.

3. A SOLUTION THROUGH COARSE DATA

The problem explained in Section 2 arises as a consequence of conditioning on data that have probability zero of being generated by the assumed sampling model, *i.e.* impossible events. The reason this occurs in statistical practice is the fundamental incompatibility between the assumed sampling model and the "observed" data: whereas the sampling model is continuous (and thus assigns zero probability to any point observation), the data always come to us in a discrete fashion, either induced by intentional rounding or grouping or through a finite precision of the measuring device.

Clearly, whenever a point value y_0 is recorded as an "observation", we do not literally believe that y_0 is the outcome of the sampling process (indeed, it can not be), but it should rather be interpreted as indicative of some (small) neighbourhood S_0 around y_0. Usual practice is to disregard this fact and simply conduct Bayesian inference on θ on the basis

of y_0 through (2.3). As explained in Section 2, however, this convenient short-cut is no longer available when $p(y_0) = \infty$. If the neighbourhood S_0 is large enough to have a non-negligible probability mass attached to it, this possibility becomes of practical relevance since the value y_0 might well be recorded as the outcome of an experiment. Clearly, as the precision of the measuring and recording scheme increases, the size of S_0 decreases and so does the probability of "observing" y_0. Whenever $p(y_0) = \infty$, inference will have to be based on the entire neighbourhood around y_0, rather than on the reported value alone. Thus, instead of (2.3) with $y = y_0$, we shall consider

$$p(\theta|y \in S_0) = \frac{P(y \in S_0|\theta)p(\theta)}{P(y \in S_0)}, \tag{3.1}$$

where $P(y \in S_0|\theta) = \int_{S_0} p(y|\theta)dy$ and $P(y \in S_0) = \int_\Theta P(y \in S_0|\theta)p(\theta)d\theta$. The crucial difference between (2.3) and (3.1) is that we now condition on an event of positive measure, namely $y \in S_0$, thus no longer contradicting the sampling assumptions. In the case of a proper prior, $p(\theta)$, this settles the issue entirely: the conditioning event has positive probability and, thus, (3.1) can immediately be used for inference on θ. If $p(\theta)$ is improper, on the other hand, we have solved the problem of conditioning on zero measure events, but we still need to check that the denominator in (3.1) is finite, so as to have a p.d.f. on θ.

The above procedure can be interpreted as follows: we are really observing a new random variable, say, $z = z(y)$ that takes values in a space, say, $\mathcal{Z}$ of subsets of $\mathcal{Y}$ that have positive probability of occurring under $p(y|\theta)$. In practice, $\mathcal{Z}$ will be a countable space. In the simplest case of directly rounding the observations, the elements of $\mathcal{Z}$ will constitute a partition of $\mathcal{Y}$. A more complicated setup is where the raw data are first rounded and afterwards transformed, which implies that the sets in $\mathcal{Z}$ are not necessarily disjoint. An example of this situation will appear in Section 5. Whenever (3.1) defines a p.d.f. for θ, the counterpart of Condition B in Section 2 applies, in the sense that we can base inference on a properly normalized distribution for θ after observing $z = S_0$. Furthermore, since $\mathcal{Z}$ is countable the conditional distribution of θ given z is defined (*i.e.* the counterpart of Condition A in Section 2 holds) if and only if $P(z = S) < \infty$ for all $S \in \mathcal{Z}$. Whereas this always obtains under a proper prior, it may fail to hold if $p(\theta)$ is an improper density function. On the other hand, as the observations will now have positive measure under the sampling model, the counterpart of A will always imply the counterpart of B. Thus, if we use a proper prior we can rely on probability theory to guarantee a properly normalized posterior distribution for every possible value z.

In practice, computing (3.1) will be more complicated than (2.3), yet quite feasible through straightforward numerical methods. In particular, we can set up the simple Gibbs sampler with the following conditionals:

$$p(\theta|y, y \in S_0) = p(\theta|y), \tag{3.2}$$

$$p(y|\theta, y \in S_0) \propto p(y|\theta)I_{S_0}(y). \tag{3.3}$$

Sequential drawing from (3.2) – (3.3) generates a Markov chain for $(y, \theta|y \in S_0)$ that will converge to the actual joint distribution and from which posterior and predictive inference

can immediately be conducted. Remark that we only require the possibility to draw from the "usual" posterior p.d.f. in (2.3) and from the sampling model, truncated to the observed set S_0. In practice, an adequate pseudo-random number generator for (3.3) will never lead to "offending" values of y for which the predictive density is not finite, since it typically operates with high precision, so that any given value y_0 has an extremely small probability of occurrence and is very unlikely to be drawn in a run of typical length.

Convergence of the Markov chain induced by (3.2) – (3.3) is always guaranteed in the practically relevant case where the support of y in the sampling does not depend on θ. The latter implies that our Gibbs sampler generates a chain on $S_0 \times \Theta$ and the Cartesian product structure assures convergence as shown in Roberts and Smith (1994). For general references in the area of Markov chain Monte Carlo and Gibbs sampling, we refer the reader to Gelfand and Smith (1990), Casella and George (1992) and Tierney (1994).

4. INDEPENDENT SAMPLING FROM SCALE MIXTURES OF NORMALS

The present Section examines a leading case where Condition A in Section 2 is fulfilled for point observations, yet Condition B does not hold for certain values of the observables.

In particular, we consider a location-scale model with errors that are independently distributed as scale mixtures of Normals. For practical purposes, the Student-t model will be the most relevant member of this class. The latter model finds increasing application in statistical practice. Maronna (1976) and Lange, Little and Taylor (1989) discuss maximum-likelihood estimation for these models, and Harvey, Ruiz and Shephard (1994) and Jacquier, Polson and Rossi (1995) use Student-t models for high-frequency financial data. Recently developed numerical methods are quite naturally adapted to the analysis of scale mixtures of Normals, in particular through the use of the Gibbs sampler under data augmentation [for the latter, see Tanner and Wong (1987)]. Details are provided in Geweke (1993) for the Student-t case and in Fernández and Steel (1996a) for general scale mixtures of Normals.

4.1. The Bayesian Model

Consider the following, frequently used, model for $y_i \in \Re$

$$y_i = \mu + \sigma\varepsilon_i,\ i = 1, \ldots, n, \tag{4.1}$$

with location parameter $\mu \in \Re$ and scale parameter $\sigma > 0$, where the ε_i's are i.i.d. scale mixtures of Normals with p.d.f.

$$p(\varepsilon_i|\nu) = \int_0^\infty f_N(\varepsilon_i|0, \lambda_i^{-1})\, dP_{\lambda_i|\nu}, \tag{4.2}$$

for some mixing probability distribution $P_{\lambda_i|\nu}$ on $\Re_+$, which can depend on a parameter $\nu \in \mathcal{N}$ of finite or infinite dimension. Leading examples, which will be studied in this Section are:

1. Normal sampling where $P_{\lambda_i|\nu}$ is a Dirac distribution on 1;
2. Finite mixtures of Normals with $P_{\lambda_i|\nu}$ a discrete distribution with finite support (with Normal sampling as a special case);
3. Student-t sampling with a Gamma$(\nu/2, \nu/2)$ mixing distribution;
4. Modulated Normal type I [see Romanowski (1979)] with Pareto$(1, \nu/2)$ mixing on the support $(1, \infty)$;
5. Modulated Normal type II [see Rogers and Tukey (1972)] where $P_{\lambda_i|\nu}$ is a Beta$(\nu/2, 1)$ distribution on $(0, 1)$.

A more extensive list of examples is provided in Fernández and Steel (1996a).

The parameters in the sampling model (4.1) – (4.2) are (μ, σ, ν), and in the prior distribution we assume the following product structure:

$$P_{(\mu,\sigma,\nu)} = P_{(\mu,\sigma)} \times P_\nu. \tag{4.3}$$

For (μ, σ) we shall adopt the commonly used improper prior with density

$$p(\mu, \sigma) \propto \sigma^{-1}, \tag{4.4}$$

which is both the Jeffreys' prior (under "independence") and the reference prior in the sense of Berger and Bernardo (1992) when ν is known [see Fernández and Steel (1995)]. The parameter of the mixing distribution ν will be assigned a probability measure P_ν.

4.2. The Analysis With Point Observations

Here we follow common statistical practice in treating the recorded observations as values $y_1, \ldots, y_n$. From the analysis in Fernández and Steel (1996a) we can derive for any mixing distribution $P_{\lambda_i|\nu}$ and any proper prior P_ν:

Result i: $p(y_1, \ldots, y_n) < \infty$ requires at least two different observations;
Result ii: if $n \geq 2$ and all observations are different, then $p(y_1, \ldots, y_n) < \infty$.

Since under a continuous sampling model the probability that any two observations are equal is zero, we can state the following result:

Theorem 1. *The Bayesian model* (4.1) – (4.4) *allows for the existence of a conditional distribution of* (μ, σ, ν) *given* $(y_1, \ldots, y_n)$ *if and only if* $n \geq 2$.

Thus, Condition A of Section 2 holds for any *any* scale mixture of Normals whenever we sample at least two observations.

Note that Result ii above does not include the zero measure event that some observations are repeated, and, thus, does not guarantee that B is fulfilled in such cases. Let us now assume that our sample contains repeated observations and let $s > 1$ denote the largest number of observations with the same value in the sample. In view of Result i, we shall always assume that $s < n$, so that the sample contains at least two different observations. With this setup, we obtain the following result:

Theorem 2. *Consider the Bayesian model* (4.1) – (4.4) *and let* s *be the largest number of observations with the same value. If* $1 < s < n$, *we obtain* $p(y_1, \ldots, y_n) < \infty$ *if and only if*

$$\int_{0<\lambda_1\leq\ldots\leq\lambda_n<\infty} \lambda_{n-s}^{-(n-2)/2} \prod_{i\neq n-s,n} \lambda_i^{1/2}\, dP_{(\lambda_1\ldots\lambda_n)} < \infty, \tag{4.5}$$

where, with a slight abuse of notation,

$$P_{(\lambda_1 \ldots \lambda_n)} = \int_{\mathcal{N}} (\prod_{i=1}^{n} P_{\lambda_i|\nu}) dP_\nu. \tag{4.6}$$

Whereas, from Theorem 1, obtaining Condition A does not depend on the particular scale mixture of Normals considered, nor on the prior P_ν, Theorem 2 implies that both intervene when we focus on Condition B.

The following theorem further examines the implications of (4.5) for the examples given in Subsection 4.1.

Theorem 3. *Under the conditions of Theorem 2, we obtain under:*
i. *Sampling from finite mixtures of Normals*: $p(y_1, \ldots, y_n) < \infty$;
ii. *Student-t or Modulated Normal type II sampling*: $p(y_1, \ldots, y_n) < \infty$ *if and only if*

$$P_\nu \left(0, \frac{s-1}{n-s}\right] = 0 \text{ and } \int_{(s-1)/(n-s)}^{\{(s-1)/(n-s)\}+\epsilon} \{(n-s)\nu - (s-1)\}^{-1} dP_\nu < \infty \text{ for all } \epsilon > 0.$$

iii. *Modulated Normal type I sampling*: $p(y_1, \ldots, y_n) < \infty$ *if and only if*

$$P_\nu \left(0, \frac{s-1}{s}\right] = 0 \text{ and } \int_{(s-1)/s}^{\{(s-1)/s\}+\epsilon} \{s\nu - (s-1)\}^{-1} dP_\nu < \infty \text{ for all } \epsilon > 0.$$

Thus, Condition B is always fulfilled when sampling from finite mixtures of Normals, and the mere fact that two observations are different suffices for inference. Interestingly, inference under Student-t or Modulated Normal sampling requires bounding ν away from zero if we wish to consider samples with repeated observations. For Modulated Normal type I sampling it is sufficient to take P_ν with support on $\nu > \{(s-1)/s\}+\epsilon$ for some $\epsilon > 0$; thus, $\nu \geq 1$ always guarantees a finite predictive value. On the other hand, under Student or Modulated Normal type II models the required lower bound for ν, $(s-1)/(n-s)$, might become as large as $n-2$.

In practice, one often chooses a prior for ν with support on all of $\Re_+$, which means that the problem will appear under Student-t or Modulated Normal sampling as soon as two observations in the sample are equal.

4.3. The Analysis With Set Observations

Let us now apply the solution proposed in Section 3 to the model $(4.1) - (4.4)$. Thus, instead of point observations, we shall consider as our data information that $y_i \in S_i, i = 1, \ldots, n$, where S_i is a neighbourhood of y_i. Since the prior assumed in $(4.3) - (4.4)$ is not proper, we need to verify whether $P(y_1 \in S_1, \ldots, y_n \in S_n) < \infty$ before inference can be conducted. The following theorem addresses this issue.

Theorem 4. *Consider the Bayesian model* $(4.1) - (4.4)$ *with any mixing distribution* $P_{\lambda_i|\nu}$ *and any proper prior* P_ν. *The observations consist of* n *intervals* $S_1, \ldots, S_n$ *(of positive*

Lebesgue measure in $\Re$*). Then* $P(y_1 \in S_1, \ldots, y_n \in S_n) < \infty$ *if and only if* $n \geq 2$ *and there exist two bounded sets, say,* S_i *and* S_j *for which*

$$\inf_{y_i \in S_i, y_j \in S_j} |y_i - y_j| > 0. \tag{4.7}$$

Thus, the existence of at least two bounded intervals that are strictly separated from each other is a necessary and sufficient condition for inference on the basis of these set observations. The necessity of this condition is the set counterpart of Result i in Subsection 4.2. Now, however, this condition is also sufficient for inference with any scale mixture of Normals. Thus, irrespective of the mixing distribution and the prior P_ν, Condition B always holds under (4.7), whereas we know that it fails for any sample not satisfying (4.7). On the other hand, Condition A will now never obtain since the collection of "offending" values, *i.e.* the samples of sets not verifying (4.7), has positive probability of being observed. Nevertheless, as stressed in Section 2, this does not preclude inference on the basis of any sample of set observations for which (4.7) holds, as is most likely in practice.

4.4. Skewed Scale Mixtures of Normals

In some situations the symmetry assumption implicit in the model (4.1) – (4.2) might be considered inappropriate for the data at hand. In such cases, we can follow the proposal of Fernández and Steel (1996b) in order to introduce skewness into the model. In particular, we can replace the density of the error term in (4.2) by

$$p(\varepsilon_i|\nu,\gamma) = \frac{2}{\gamma + \frac{1}{\gamma}} \int_0^\infty \{f_N(\varepsilon_i/\gamma|0,\lambda_i^{-1}) I_{[0,\infty)}(\varepsilon_i) + f_N(\gamma\varepsilon_i|0,\lambda_i^{-1}) I_{(-\infty,0)}(\varepsilon_i)\} dP_{\lambda_i|\nu}, \tag{4.8}$$

where ν is as before and we introduce a parameter $\gamma \in \Re_+$. Thus, (4.8) is obtained from (4.2) by scaling with γ to the right of the origin and with its inverse to the left of zero. Clearly, for $\gamma = 1$ (4.8) coincides with (4.2), but if $\gamma > 1$ we introduce right skewness, whereas values of $\gamma < 1$ lead to left skewed distributions. More details on the properties of such distributions are provided in Fernández and Steel (1996b).

The prior distribution is now given by

$$P_{(\mu,\sigma,\nu,\gamma)} = P_{(\mu,\sigma)} \times P_\nu \times P_\gamma, \tag{4.9}$$

where $P_{(\mu,\sigma)}$ is described in (4.4) and P_ν and P_γ are any probability measures.

The following result addresses the influence of our skewness transformation.

Theorem 5. *Consider the Bayesian model* (4.1), (4.4), (4.8) – (4.9).

i. With point observations $y_1, \ldots, y_n$ *we obtain* $p(y_1, \ldots, y_n) < \infty$ *if and only if the same holds when* $\gamma = 1$.

ii. With set observations $S_1, \ldots, S_n$*, we obtain* $P(y_1 \in S_1, \ldots, y_n \in S_n) < \infty$ *if and only if the same holds when* $\gamma = 1$.

Surprisingly, the extra flexibility in dealing with skewness does not affect the possibility of conducting inference, although the actual numerical results might, of course, be

quite different. From Theorem 5 we can, in fact, conclude that both Condition A and B hold in exactly the same circumstances as for the symmetric model. Thus, all results presented in Subsections 4.2 and 4.3 immediately apply to the skewed case.

A convenient algorithm, based on the Gibbs sampler, for the numerical analysis of skewed Student-t models was presented in Fernández and Steel (1996b) for the case of point observations. If we wish to conduct inference from set observations, we merely need to add the conditional in (3.3) to the Gibbs sampler. The next Section will present an application of skewed Student sampling to a financial data set.

5. AN APPLICATION TO STOCK PRICE RETURNS

5.1. The Model and the Data

The data we will examine here were taken from Buckle (1995), and represent a sample of 49 returns on Abbey National shares between July 31 and October 8, 1991. These returns are constructed from price data p_i, $i = 0, \ldots, 49$, as $y_i = (p_i - p_{i-1})/p_{i-1}$, $i = 1, \ldots, 49$. As the data seem to exhibit some skewness, Buckle (1995) uses a Stable distribution, allowing for asymmetry.

Here, we shall follow Fernández and Steel (1996b) and use instead the skewed Student sampling model obtained from (4.1) and (4.8) with $P_{\lambda_i|\nu}$ a Gamma$(\nu/2, \nu/2)$ distribution. This leads to the following sampling density:

$$\begin{aligned} p(y_i|\mu,\sigma,\nu,\gamma) =& 2\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{1/2}}\frac{1}{\sigma\left(\gamma+\frac{1}{\gamma}\right)} \\ & \left[1+\frac{(y_i-\mu)^2}{\nu\sigma^2}\left\{\frac{1}{\gamma^2}I_{[0,\infty)}(y_i-\mu)+\gamma^2 I_{(-\infty,0)}(y_i-\mu)\right\}\right]^{-\frac{\nu+1}{2}}, \end{aligned} \tag{5.1}$$

which we combine with the prior distribution in (4.9) where $P_{(\mu,\sigma)}$ is as described in (4.4). In this particular application, we shall choose an exponential prior distribution for ν with mean 10 and variance 100, spreading the prior mass over a wide range of tail behaviour, and a Normal$(0, \pi/2)$ distribution truncated to $\Re_+$ for γ. The latter centers the prior over $\gamma = 1$, *i.e.* symmetry, and provides a compromise between sufficient spread and reasonably equal prior weights to right and left skewness. Fernández and Steel (1996b) provides more details on prior elicitation.

Let us first consider the analysis with point observations: Theorem 1 assures us that Condition A in Section 2 holds as $n \geq 2$. However, the data contain seven observations that are recorded as zero. Thus, from Theorem 3 (ii) we know that Condition B does not hold with this data set, since $(s-1)/(n-s) = 6/42 = 1/7$ and the prior distribution for ν has mass arbitrarily close to zero. Bayesian inference on the basis of this sample is, therefore, precluded. This problem was avoided in Fernández and Steel (1996b) by slightly perturbing the original data, thus avoiding repeated observations. However, this solution is arbitrary and not in accordance with the way the data are recorded. Here we will, instead, consider the solution proposed in Section 3.

The set observations corresponding to this sample are constructed as follows: prices were recorded in integer values (in Pence) and we shall assume they were rounded to the nearest integer. The set observations for the returns are then defined as

$$S_i = \left(\frac{p_i - p_{i-1} - 1}{p_{i-1} + 0.5}, \frac{p_i - p_{i-1} + 1}{p_{i-1} - 0.5} \right), \tag{5.2}$$

$i = 1, \ldots, 49$. As a consequence of the return transformation after rounding the prices, the sets S_i are not all pairwise disjoint, yet we can find at least two sets for which (4.7) holds. Thus, Bayesian inference on the basis of set observations is possible from Theorem 4. Figure 1 graphically displays the data information as follows: the set observations are located on the horizontal axis and each has mass $1/n$ assigned to it in a Uniform way. Density values are plotted on the vertical axis. Whenever two or more sets intersect, the intersection area gets assigned the sum of the density values. In other words, the figure displays a mixture of Uniform density functions on each of the set observations with weights equal to $1/n$. Some evidence of right skewness seems apparent from this plot.

5.2. Numerical Results

The numerical analysis will be conducted as indicated in Section 3. In this particular model, data augmentation with the mixing parameters $\lambda_1, \ldots, \lambda_n$ will facilitate the Gibbs sampler used for the posterior analysis. Thus, the complete Gibbs sampler will be conducted on $(y_1, \ldots, y_n, \mu, \sigma, \nu, \gamma, \lambda_1, \ldots, \lambda_n)$. For the full conditionals of μ, σ, ν, γ and $(\lambda_1, \ldots, \lambda_n)$ we refer the reader to Fernández and Steel (1996b). Whereas the latter constitutes (3.2), we now need to add the full conditional distribution of $(y_1, \ldots, y_n)$ [*i.e.* (3.3)], which is given by the product of the p.d.f.'s

$$\begin{aligned} &p(y_i|\mu, \sigma, \nu, \gamma, \lambda_i, y_i \in S_i) \propto \\ &\quad \exp\left[-\frac{\lambda_i}{2\sigma^2}(y_i - \mu)^2 \left\{ \frac{1}{\gamma^2} I_{[0,\infty)}(y_i - \mu) + \gamma^2 I_{(-\infty,0)}(y_i - \mu) \right\} \right] I_{S_i}(y_i), \end{aligned} \tag{5.3}$$

for $i = 1, \ldots, n$. Thus, given all the rest, the y_i's are independent random variables with a skewed Normal distribution truncated to the corresponding set observation S_i. Drawings from a truncated skewed Normal distribution are generated as explained in Appendix B. In all, the Gibbs sampler generates a Markov chain in $2n + 4$ dimensions by cycling through six steps. Predictive inference will be conducted through averaging the sampling density in (5.1), following the Rao-Blackwell argument suggested in Gelfand and Smith (1990).

The continuous lines in Figures 2-5 display the posterior p.d.f.'s of $\mu, \tau = \sigma^{-1}, \gamma$ and ν for the set observations in (5.2), based on a sequential Gibbs run of 250,000 drawings, after discarding the initial 10,000 values (the "burn-in"). As expected, some evidence for right skewness transpires from Figure 4 as values for $\gamma > 1$ receive most of the posterior mass. The data also indicate some support for relatively thick tails (Figure 5), although the small data set under consideration is not very informative on tail behaviour. Figure 1 shows the predictive density function, overplotting the data information. Clearly, the predictive distribution fits the data quite well.

We contrast this analysis with the one based on perturbed point observations, using the same number of drawings in the Gibbs sampler. The perturbation was applied to the price data, p_i, and consisted in adding a Uniformly distributed random number on $(-5\times10^{-7}, 5\times10^{-7})$ to the recorded prices, who are themselves of the order 300. For a given perturbation, the resulting point observations $y_i = (p_i - p_{i-1})/p_{i-1}$, $i = 1, \ldots, 49$ no longer contained any repeated values and dashed lines in Figures 1-5 summarize posterior and predictive inference. With regards to the predictive distribution, the following technical issue arises (under point observations with a prior P_ν that has mass arbitrarily close to zero): From Theorems 3 (ii) and 5 (i), whenever two observations are equal the within-sample predictive density value is infinite; this implies that the out-of-sample p.d.f. is also infinite at each observed data point. In addition, we can prove that the latter p.d.f. is unbounded in a neighbourhood of each observation. However, such neighbourhoods contain a negligible amount of probability mass and, therefore, the smooth (dashed) curve depicted in Figure 1 is a very good approximation to the actual predictive p.d.f. for all practical purposes.

Even though the results on the basis of perturbed point observations are very close to those with set observations, the particular perturbation employed is, of course, arbitrary. For that reason, we would hesitate to implement a more substantial perturbation on such an ad-hoc basis. As indicated by the present empirical evidence, the choice between set observations and a small ad-hoc perturbation need not be a major issue in cases where the problematic area receives very little posterior mass. We remind the reader that problems occur for the original unperturbed point observations whenever $\nu \leq 1/7$. As Figure 5 shows, very little posterior probability is allocated to that region for ν. Since the Markov chain is unlikely to wander in this area, the particular solution adopted need not make a large difference in this case. If we force the issue, however, and fix ν at a problematic value, say $\nu = 0.1 < 1/7$, we observe a very different picture.

5.3. Fixing the Degrees of Freedom

Clearly, the tails of the Student-t sampling model with $\nu = 0.1$ are far too thick to adequately fit this data set, which displays quite a concentration of mass around the mode (see Figure 1). As a consequence, the model will try to accommodate the empirical mass around the mode by increasing the precision $\tau = \sigma^{-1}$. Thus, the observations that are not close to the mode will tend to be regarded as "outliers" with relatively small weights (*i.e.* small values of the mixing variable λ_i) attached to them. This happens both when set observations are used and with perturbed data. However, the degree to which this phenomenon affects the results is quite different.

Figures 6-8 graphically display the posterior p.d.f.'s of μ, $\ln(\tau)$ and γ, whereas Figure 9 graphs the natural logarithm of the predictive density function. Let us first comment on the results using set observations; as expected, the precision, τ, has its mass at much higher values than in the case with free ν [note we now graph $\ln(\tau)$]. As the resulting predictive distribution is very spiked (see Figure 9), it will essentially choose μ so as to best fit the mode of the data. As is clear from Figure 1, this mode is not unequivocally determined, and, as a result, the posterior of μ will switch between the local modes in the empirical distribution, depending on how $(y_1, \ldots, y_n)$ are drawn in their intervals $(S_1, \ldots, S_n)$. This

strange behaviour of μ should be a clear warning to the practitioner that the model with $\nu = 0.1$ is not a good choice for this data set. The inference on the skewness parameter, γ, is surprisingly little affected by the restriction on ν.

If we use perturbed point observations, the concentration of the data around zero is much higher: whereas the seven repeated observations roughly lie in the set $(-0.033, 0.033)$ if we use set observations, the corresponding perturbed point observations are all situated in the interval $(-2\times10^{-9}, 2\times10^{-9})$. This translates into a much higher precision, evident from Figure 7. Virtually all the weight is now assigned to the seven perturbed zero observations (λ_i's of the order 10), whereas the 42 remaining observations are practically discarded (λ_i's of the order 10^{-11}). As a consequence, μ gets almost all of its mass very close to zero (posterior mean is 6×10^{-11} and the standard deviation is 6×10^{-9}) and evidence on the right skewness in the data is now lost (Figures 6 and 8). Indeed, the predictive distribution graphed in Figure 9 is even slighty left skewed, as a result of the particular distribution of the perturbed zero observations (it so happens that they are somewhat bunched on the negative axis, and five out of the seven are situated to the left of the posterior mean on μ). Note that, due to the extremely spiked shape of the predictive, we have plotted the logarithm of the density value, rendering graphical comparison with the predictive from set observations possible.

Especially in less extreme cases than this one, the results from the model with perturbed point observations might well lead to the mistaken impression that the model fits the data well since precision is high and posterior distributions are quite concentrated. The Gibbs sampler for this model is very slow to converge, so all the results in this Subsection were based on a chain of 150,000 drawings after a burn-in of 150,000 drawings. Even though convergence proved much faster for the model using set observations, the same setup was used there. Finally, in terms of speed of execution, there is no important difference between set observations and point observations, since Gauss-386i VM (Version 3.2.13) programs for the latter executed at a rate of approximately 35,000 drawings per hour on a Pentium 90 CPU, whereas this was 25,000 per hour for the analysis through set observations. In the more relevant case with unrestricted degrees of freedom, ν, these numbers are 31,000 and 23,000, respectively.

Clearly, when we move to more dangerous waters by imposing ν equal to a value for which the original point observations do not allow for inference, the issue of how this problem is resolved becomes of critical importance. We run into problems if we use small ad-hoc perturbations, whereas larger perturbations risk seriously biasing the inference. The only real solution to the problem seems, in our view, to be through a coherent use of set observations.

6. CONCLUDING REMARKS

In this paper, we have identified a potential problem arising from a fundamental incompatibility between the recorded point observations and a continuous sampling model. In a Bayesian context, this problem can preclude inference even under a proper prior distribution. As a consequence of rounding, this phenomenon becomes of real practical

importance.

The solution we propose is based on the way the data were actually recorded and considers set observations, *i.e.* intervals around the recorded point observations. Once we formally conduct the analysis conditionally upon samples of set observations, we can accommodate continuous sampling assumptions since such samples have nonzero measure. We implement this solution through a Gibbs sampler, which cycles through the sampling model, given the set observations, and the "usual" posterior distribution with point observations.

Our leading example is that of scale mixtures of Normals, in the context of a location-scale model under a "usual" noninformative prior. The case of independent sampling from a possibly skewed Student-t distribution is of prime empirical importance. Problems essentially revolve around the presence of repeated observations and we characterize those samples of point observations for which inference is precluded. For finite mixtures of Normals this lack of inference only occurs if all the observations are the same, but *e.g.* for Student-t sampling with unrestricted degrees of freedom it appears whenever two observations in the sample are recorded as having the same value.

In a numerical example with skewed Student-t sampling, we contrast the formal analysis based on set observations with an ad-hoc procedure consisting in slightly perturbing the data, to avoid the exact repetition of point observations. We show that a small perturbation can lead to senseless and misleading results in cases where problematic values of the degrees of freedom parameter have empirical posterior support. Clearly, larger ad-hoc perturbations would risk biasing the inference in one direction or another. In addition, the numerical implementation using set observations is only marginally more computationally intensive. Thus, the analysis through set observations seems a much preferred solution to this problem.

ACKNOWLEDGEMENTS

We gratefully acknowledge stimulating discussions with Michael Lavine, Jean-Francois Richard and Richard Smith, as well as useful comments from an anonymous referee. Part of this research was carried out at the Statistics Department of Purdue University, facilitated by a travel grant of the Netherlands Organization for Scientific Research (NWO).

APPENDIX A: Proofs of the Theorems

Proof of Theorem 2

The Bayesian model in (4.1) – (4.4) leads to

$$p(y_1,\ldots,y_n) \propto \int_{\Re\times\Re_+\times\Re_+^n} \Big(\prod_{i=1}^n \lambda_i^{1/2}\Big)\sigma^{-(n+1)} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n \lambda_i(y_i-\mu)^2\right\} d\mu d\sigma dP_{(\lambda_1,\ldots,\lambda_n)}, \tag{A.1}$$

with $P_{(\lambda_1,\ldots,\lambda_n)}$ as defined in (4.6). After integrating out μ with a Normal distribution and σ through a Gamma distribution on σ^{-2} we are left with

$$p(y_1,\ldots,y_n) \propto \int_{\Re_+^n} \left(\prod_{i=1}^n \lambda_i^{1/2}\right) \left(\sum_{i=1}^n \lambda_i\right)^{(n-2)/2} S^2(\lambda,y)^{-(n-1)/2} dP_{(\lambda_1,\ldots,\lambda_n)}, \qquad (A.2)$$

where

$$S^2(\lambda,y) = \sum_{1\le i<j\le n} \lambda_i\lambda_j(y_i-y_j)^2. \qquad (A.3)$$

Note that $\sum_{i=1}^n \lambda_i$ has upper and lower bounds which are both proportional to the biggest λ_i, whereas $S^2(\lambda,y)$ has upper and lower bounds proportional to the biggest product $\lambda_i\lambda_j$ for which $y_i \neq y_j$. Since the largest number of observations with the same value is s, Theorem 2 follows.

Proof of Theorem 3

From Theorem 2, we need to check whether (4.5) is fulfilled for each of the sampling distributions considered.

i. *Sampling from finite mixtures of Normals:*

This corresponds to $P_{(\lambda_1,\ldots,\lambda_n)} = \prod_{i=1}^n P_{\lambda_i}$, where P_{λ_i} is a discrete distribution with finite support. Thus (4.5) always holds for any value of $s = 2,\ldots,n-1$.

ii. *Student-t sampling:*

In this case, $P_{\lambda_i|\nu}$ is a Gamma$(\nu/2,\nu/2)$ distribution [the p.d.f. of which is denoted by $f_G(\lambda_i|\nu/2,\nu/2)$] and, by Fubini's theorem, (4.5) can be computed as

$$\int_0^\infty I(\nu) dP_\nu, \qquad (A.4)$$

where

$$I(\nu) = \int_{0<\lambda_1\le\ldots\le\lambda_n<\infty} \lambda_{n-s}^{-(n-2)/2} \prod_{i\neq n-s,n} \lambda_i^{1/2} \prod_{i=1}^n f_G\left(\lambda_i|\frac{\nu}{2},\frac{\nu}{2}\right) d\lambda_1\ldots d\lambda_n. \qquad (A.5)$$

In what follows, we shall make use of the bounds

$$\frac{w^v}{v}\exp(-rw) \le \int_0^w \lambda^{v-1}\exp(-r\lambda)d\lambda \le \frac{w^v}{v}, \text{ for any } r,v,w>0. \qquad (A.6)$$

Necessity:

Iterative use of the lower bound in $(A.6)$ while integrating $\lambda_1,\ldots,\lambda_n$ shows that $I(\nu)<\infty$ requires $\nu > (s-1)/(n-s)$. When the latter holds, we obtain a lower bound for $I(\nu)$ proportional to

$$H_1(\nu) = n^{-n\nu/2}\left\{\Gamma\left(\frac{\nu}{2}\right)\right\}^{-n}\Gamma\left(\frac{n\nu}{2}\right)\frac{1}{(\nu+1)^{n-s-1}\prod_{l=1}^s\{(n-l)\nu-(l-1)\}}, \qquad (A.7)$$

which, in turn, has a lower bound proportional to $\{(n-s)\nu-(s-1)\}^{-1}$ in any region where $(s-1)/(n-s)<\nu<\{(s-1)/(n-s)\}+\epsilon$ for some $\epsilon>0$. Thus, the necessity of the conditions in Theorem 3 (ii) follows.

Sufficiency:

Assuming that $\nu>(s-1)/(n-s)$ and now applying the upper bound in $(A.6)$, we obtain an upper bound for $I(\nu)$ proportional to $H_2(\nu)=n^{n\nu/2}H_1(\nu)$, with $H_1(\nu)$ as in $(A.7)$. When $(s-1)/(n-s)<\nu<\{(s-1)/(n-s)\}+\epsilon$ for some $\epsilon>0$, $H_2(\nu)$ is bounded from above by a constant times $\{(n-s)\nu-(s-1)\}^{-1}$.

Finally, in order to study the behaviour of $I(\nu)$ as $\nu\to\infty$ we use the fact that, for $\lambda_1\le\ldots\le\lambda_n$, $\prod_{i\neq n-s,n}\lambda_i^{1/2}\le\lambda_{n-s}^{(n-s-1)/2}\lambda_n^{(s-1)/2}$. We then have

$$\begin{aligned} I(\nu) &\le \int_0^\infty \lambda_{n-s}^{-(s-1)/2} f_G\left(\lambda_{n-s}|\frac{\nu}{2},\frac{\nu}{2}\right)d\lambda_{n-s}\int_0^\infty \lambda_n^{(s-1)/2} f_G\left(\lambda_n|\frac{\nu}{2},\frac{\nu}{2}\right)d\lambda_n \\ &= \Gamma\left(\frac{\nu-(s-1)}{2}\right)\Gamma\left(\frac{\nu+(s-1)}{2}\right)\left\{\Gamma\left(\frac{\nu}{2}\right)\right\}^{-2}. \end{aligned} \tag{A.8}$$

Since for all z bigger than a positive constant $\Gamma(z)$ has upper and lower bounds proportional to $z^{z-(1/2)}\exp(-z)$ [see Whittaker and Watson (1927, chap.12)], the expression in $(A.8)$ has a finite limit as $\nu\to\infty$. This establishes the sufficiency of the conditions stated in Theorem 3 (ii).

Modulated Normal type II sampling:

We again compute (4.5) following $(A.4)-(A.5)$, replacing $f_G(\lambda_i|\nu/2,\nu/2)$ by $f_B(\lambda_i|\nu/2,1)$, the p.d.f. of a Beta distribution. Direct calculations show that $I(\nu)<\infty$ requires $\nu>(s-1)/(n-s)$, in which case

$$I(\nu)\propto\frac{\nu^{n-1}}{(\nu+1)^{n-s-1}\prod_{l=1}^{s}\{(n-l)\nu-(l-1)\}}. \tag{A.9}$$

The latter expression defines a continuous function of $\nu>(s-1)/(n-s)$, with a finite limit as $\nu\to\infty$. On the other hand, when $(s-1)/(n-s)<\nu<\{(s-1)/(n-s)\}+\epsilon$ for some $\epsilon>0$, this expression has upper and lower bounds proportional to $\{(n-s)\nu-(s-1)\}^{-1}$. This immediately leads to Theorem 3 (ii) under this sampling model.

iii. *Modulated Normal type I sampling:*

We again compute (4.5) from $(A.4)-(A.5)$, now replacing the Gamma p.d.f. by $f_P(\lambda_i|1,\nu/2)$, which corresponds to a Pareto distribution (of the first kind) with support on $(1,\infty)$. Integrating out the variables in the order $\lambda_n,\ldots,\lambda_1$ immediately shows that $I(\nu)<\infty$ requires $\nu>(s-1)/s$ and, provided that this holds,

$$I(\nu)\propto\frac{\nu^{n}}{\prod_{l=1}^{s}\{l\nu-(l-1)\}\prod_{l=s+1}^{n}\{l\nu+(n-l)\}}. \tag{A.10}$$

Thus, $I(\nu)$ is a continuous function of $\nu>(s-1)/s$, with a finite limit as $\nu\to\infty$ and with upper and lower bounds proportional to $\{s\nu-(s-1)\}^{-1}$ if $(s-1)/s<\nu<\{(s-1)/s\}+\epsilon$ with $\epsilon>0$. Theorem 3 (iii) therefore obtains.

Proof of Theorem 4

After integrating out μ and σ from $(A.1)$, which requires $n \geq 2$, we are left with $p(y_1, \ldots, y_n)$ in $(A.2)$, which still needs to be integrated over the sets $S_1, \ldots, S_n$. Applying Fubini's theorem, we shall first perform the integral over these sets, dealing with the integral with respect to $P_{(\lambda_1,\ldots,\lambda_n)}$ afterwards. Thus, we are first concerned with evaluating

$$T(\lambda) = \int_{S_1 \times \ldots \times S_n} S^2(\lambda, y)^{-(n-1)/2} dy_1 \ldots dy_n, \tag{A.11}$$

with $S^2(\lambda, y)$ defined in $(A.3)$.

<u>*Sufficiency:*</u>

Let us assume that (4.7) holds for, say, S_1 and S_2. First we observe that

$$S^2(\lambda, y) = \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} \big(\sum_{i=1}^{n} \lambda_i\big) \eta_2^2 + (\eta_3 - \rho, \ldots, \eta_n - \rho) Q (\eta_3 - \rho, \ldots, \eta_n - \rho)', \tag{A.12}$$

where $\eta_i = y_1 - y_i$ for $i = 2, \ldots, n$, $\rho = \lambda_2 \eta_2 / (\lambda_1 + \lambda_2)$ and $Q = (q_{ij})_{i,j=3}^{n}$ is an $(n-2) \times (n-2)$ positive definite symmetric (PDS) matrix with diagonal elements given by $q_{ii} = \lambda_i \sum_{j \neq i} \lambda_j$, whereas the off-diagonal elements are $q_{ij} = q_{ji} = -\lambda_i \lambda_j$.

Since, by assumption, $|\eta_2| \geq K$ for some constant $K > 0$, $(A.12)$ implies that the integrand in $(A.11)$ is the kernel of an $(n-2)$-variate Cauchy distribution for $(\eta_3, \ldots, \eta_n)'$. Making a transformation from $y_1, \ldots, y_n$ to $y_1, \eta_2, \ldots, \eta_n$ and integrating $(\eta_3, \ldots, \eta_n)'$ over the whole of $\Re^{n-2}$ using the latter Cauchy distribution, leads to

$$T(\lambda) \leq \big(\prod_{i=1}^{n} \lambda_i^{-1/2}\big) \big(\sum_{i=1}^{n} \lambda_i\big)^{-(n-2)/2} \int_{\{y_1 \in S_1, y_1 - \eta_2 \in S_2\}} |\eta_2|^{-1} dy_1 d\eta_2. \tag{A.13}$$

The integral in $(A.13)$ is finite since S_1 and S_2 are bounded and $|\eta_2| \geq K > 0$. Combining $(A.2)$, $(A.11)$ and $(A.13)$ immediately implies that $P(y_1 \in S_1, \ldots, y_n \in S_n) < \infty$ under any probability measure $P_{(\lambda_1,\ldots,\lambda_n)}$.

<u>*Necessity:*</u>

Defining $\eta_2, \ldots, \eta_n$ as before, we have $S^2(\lambda, y) = \eta' \tilde{Q} \eta$, where $\eta = (\eta_2, \ldots, \eta_n)'$ and $\tilde{Q} = (q_{ij})_{i,j=2}^{n}$ with the elements q_{ij} defined in the same way as the elements of Q in $(A.12)$. Since $\tilde{Q}$ is a PDS matrix, it can be expressed by the Schur decomposition theorem as $\tilde{Q} = O' D O$, for an orthogonal matrix O and a diagonal matrix $D = diag(d_2, \ldots, d_n)$ whose diagonal elements are the eigenvalues of $\tilde{Q}$.

We consider a variable transformation from $y_1, \ldots, y_n$ to $y_1, \xi_2, \ldots, \xi_n$, where $\xi = (\xi_2, \ldots, \xi_n)' = O\eta = O(y_1 - y_2, \ldots, y_1 - y_n)'$. Since (4.7) is assumed not to hold, the image set of $S_1 \times \ldots \times S_n$ in the transformed variables will contain an $(n-1)$-dimensional connected set, C, for $(\xi_2, \ldots, \xi_n)'$, the closure of which contains the $(n-1)$-dimensional vector of zeros. This leads to

$$\begin{aligned} T(\lambda) &\geq \int_{S_1} dy_1 \int_C \big(\sum_{i=2}^{n} d_i \xi_i^2\big)^{-(n-1)/2} d\xi_2 \ldots d\xi_n \\ &\geq \Big(\max_{i=1,\ldots,n} d_i\Big)^{-(n-1)/2} \int_{S_1} dy_1 \int_C \big(\sum_{i=2}^{n} \xi_i^2\big)^{-(n-1)/2} d\xi_2 \ldots d\xi_n. \end{aligned}$$

The last integral is seen to be infinite after a polar transformation.

Proof of Theorem 5

From the unimodality of the Normal distribution, the following upper and lower bounds for $p(\varepsilon_i|\nu,\gamma)$ in (4.8) can be derived

$$\frac{2}{\gamma+\frac{1}{\gamma}}\int_0^\infty f_N\left(\frac{\varepsilon_i}{h(\gamma)}|0,\lambda_i^{-1}\right)dP_{\lambda_i|\nu}, \tag{A.15}$$

with

$$h(\gamma)=\begin{cases}\max\{\gamma,1/\gamma\} & \text{for the upper bound}\\ \min\{\gamma,1/\gamma\} & \text{for the lower bound.}\end{cases} \tag{A.16}$$

This allows us to bound $p(y_1,\ldots,y_n)$ by

$$2^n\int_{\Re\times\Re_+\times\Re_+\times\Re_+^n}(\prod_{i=1}^n\lambda_i^{1/2})\frac{\sigma^{-(n+1)}}{(\gamma+\frac{1}{\gamma})^n}\exp\left\{-\frac{1}{2\sigma^2h(\gamma)^2}\sum_{i=1}^n\lambda_i(y_i-\mu)^2\right\}d\mu d\sigma dP_\gamma dP_{(\lambda_1,\ldots,\lambda_n)}, \tag{A.17}$$

with $h(\gamma)$ as in (A.16). After transforming from σ to $\vartheta=h(\gamma)\sigma$, the expression in (A.17) is seen to be equal to

$$2^n\int_0^\infty\left(\frac{h(\gamma)}{\gamma+\frac{1}{\gamma}}\right)^n dP_\gamma\int_{\Re\times\Re_+\times\Re_+^n}(\prod_{i=1}^n\lambda_i^{1/2})\vartheta^{-(n+1)}\exp\left\{-\frac{1}{2\vartheta^2}\sum_{i=1}^n\lambda_i(y_i-\mu)^2\right\}d\mu d\vartheta dP_{(\lambda_1,\ldots,\lambda_n)}. \tag{A.18}$$

Clearly, for both choices of $h(\gamma)$ in (A.16), the value of the first integral in (A.18) lies in the interval $(0,1)$ under any proper prior P_γ. On the other hand, the second integral in (A.18) corresponds to $p(y_1,\ldots,y_n)$ when γ is fixed at the value 1. This proves Theorem 5.

APPENDIX B: Drawing From a Truncated Skewed Normal Distribution

In order to draw from (5.3), we distinguish three possible situations, depending on the location of the set S_i in (5.2) relative to μ.

i. If $\inf S_i \geq \mu$, the relevant interval is entirely to the right of the mode, and we can simply draw from a Normal$(\mu,\gamma^2\sigma^2/\lambda_i)$ distribution for y_i, truncated to S_i. Drawings from a truncated Normal distribution are generated through the mixed rejection algorithm described in Geweke (1991);

ii. If $\sup S_i \leq \mu$, y_i is drawn from a Normal$(\mu,\sigma^2/(\gamma^2\lambda_i))$ distribution truncated to S_i;

iii. If $\inf S_i < \mu < \sup S_i$, we need to take into account that the shape of the distribution varies within the interval S_i, as it now extends both sides of the mode. Denoting by $\Phi(\cdot)$ the cumulative distribution function of the standard Normal distribution, it can easily be

shown that the probability of y_i being to the right of the mode is:

$$P(y_i \geq \mu | \mu, \sigma, \nu, \gamma, \lambda_i, y_i \in S_i) = \frac{\gamma \left[\Phi \left\{ \lambda_i^{1/2} \gamma^{-1} \sigma^{-1} (\sup S_i - \mu) \right\} - \frac{1}{2} \right]}{\gamma \left[\Phi \left\{ \lambda_i^{1/2} \gamma^{-1} \sigma^{-1} (\sup S_i - \mu) \right\} - \frac{1}{2} \right] + \frac{1}{\gamma} \left[\frac{1}{2} - \Phi \left\{ \lambda_i^{1/2} \gamma \sigma^{-1} (\inf S_i - \mu) \right\} \right]}. \quad (B.1)$$

So, with the probability in $(B.1)$, we draw y_i from a Normal$(\mu, \gamma^2\sigma^2/\lambda_i)$ distribution truncated to $[\mu, \sup S_i)$ and with the complementary probability y_i is drawn from a Normal$(\mu, \sigma^2/(\gamma^2\lambda_i))$ distribution intersected with the interval $(\inf S_i, \mu)$.

REFERENCES

Ball, C.A. (1988) Estimation bias induced by discrete security prices. *J. Finance*, **43**, 841-865.

Berger, J.O., and Bernardo, J.M. (1992) On the development of reference priors (with discussion). In *Bayesian Statistics 4* (eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith), pp. 35-60. Oxford: Oxford University Press.

Buckle, D.J. (1995) Bayesian inference for stable distributions. *J. Am. Statist. Ass.*, **90**, 605-613.

Casella, G. and George, E. (1992). Explaining the Gibbs sampler. *Am. Statistn*, **46**, 167-174.

DeGroot, M.H. (1970) *Optimal Statistical Decisions*. New York: McGraw-Hill.

Dempster, A.P. and Rubin, D.B. (1983) Rounding error in regression: The appropriateness of Sheppard's corrections. *J. R. Statist. Soc.* B, **45**, 51-59.

Fernández, C., and Steel, M.F.J. (1995) Reference priors in non-Normal location problems. *Discussion Paper 9591*. CentER, Tilburg University, The Netherlands.

Fernández, C., and Steel, M.F.J. (1996a) On Bayesian inference under sampling from scale mixtures of Normals. *Discussion Paper 9602*. CentER, Tilburg University, The Netherlands.

Fernández, C., and Steel, M.F.J. (1996b) On Bayesian modelling of fat tails and skewness. *Discussion Paper 9658*. CentER, Tilburg University, The Netherlands.

Fisher, F.A. (1922) On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc.* A, **222**, 309-368.

Florens, J.P., Mouchart, M. and Rolin, J.M. (1990) Invariance arguments in Bayesian statistics. In *Economic Decision Making: Games, Econometrics and Optimisation* (eds. J. Gabszewicz, J.F. Richard and L.A. Wolsey). Amsterdam: North-Holland.

Gelfand, A., and Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398-409.

Geweke, J. (1991) Efficient simulation from the multivariate Normal and Student-t distributions subject to linear constraints. In *Computing Science and Statistics* (eds. E.M. Keramidas and S.M. Kaufman), pp. 571-578. Fairfax Station, VA: Interface Foundation.

Geweke, J. (1993) Bayesian treatment of the independent Student-t linear model. *J. Appl. Econometr.*, **8**, S19-S40.

Harvey, A.C., Ruiz, E. and Shephard, N.G. (1994) Multivariate stochastic variance models. *Rev. Economic Stud.*, **61**, 247-264.

Hausman, J.A., Lo, A.W. and MacKinlay, A.C. (1992) An ordered probit analysis of transaction stock prices. *J. Fin. Economics*, **31**, 319-379.

Heitjan, D.F. (1989) Inference from grouped continuous data: A review. *Stat. Sci.*, **4**, 164-183.

Heitjan, D.F., and Rubin, D.B. (1991) Ignorability and coarse data. *Ann. Statist.*, **19**, 2244-2253.

Jacquier, E., Polson, N.G. and Rossi, P.E. (1995) Stochastic volatility: Univariate and multivariate extensions. *Mimeo.* Graduate School of Business, University of Chicago.

Lange, K.L., Little, R.J.A. and Taylor, J.M.G. (1989) Robust statistical modeling using the t-distribution. *J. Am. Statist. Ass.*, **84**, 881-896.

Lindley, D.V. (1950) Grouping corrections and maximum likelihood equations. *Proc. Camb. Phil. Soc.*, **46**, 106-110.

Maronna, R. (1976) Robust M-estimators of multivariate location and scatter. *Ann. Statist.*, **4**, 51-67.

Mouchart, M. (1976) A note on Bayes theorem. *Statistica*, **36**, 349-357.

Roberts, G.O., and Smith, A.F.M. (1994) Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stoch. Processes Applicat.*, **49**, 207-216.

Rogers, W.H. and Tukey, J.W. (1972) Understanding some long-tailed symmetric distributions. *Statistica Neerlandica*, **26**, 211-226.

Romanowski, M. (1979) *Random Errors in Observation and the Influence of Modulation on Their Distribution.* Stuttgart: Verlag Konrad Wittwer.

Sheppard, W.F. (1898) On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale. *Proc. London Math. Soc.*, **29**, 353-380.

Tanner, M.A., and Wong, W.H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Statist. Ass.*, **82**, 528-550.

Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.*, **22**, 1701-1762.

Whittaker, E.T. and Watson, G.N. (1927) *A Course of Modern Analysis: An Introduction to the General Theory of Infinite Processes and of Analytic Functions.* Cambridge: Cambridge University Press.

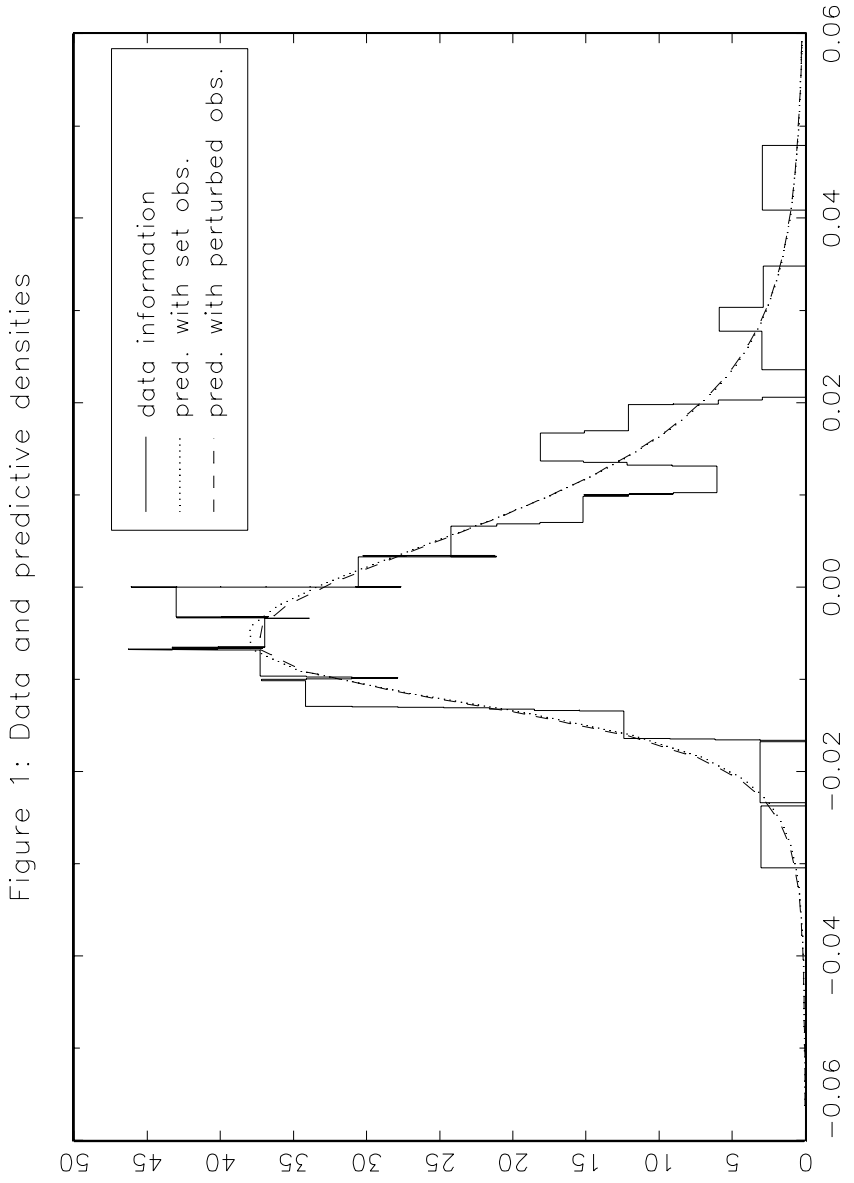

Figure 1: Data and predictive densities

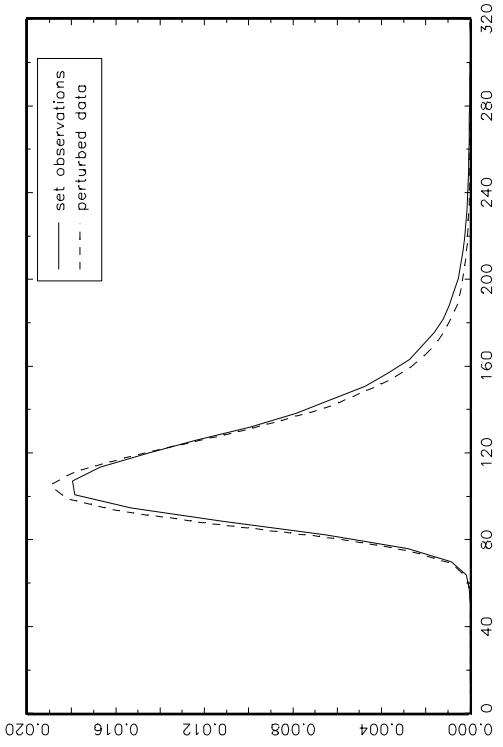

Figure 3: Posterior Density for tau

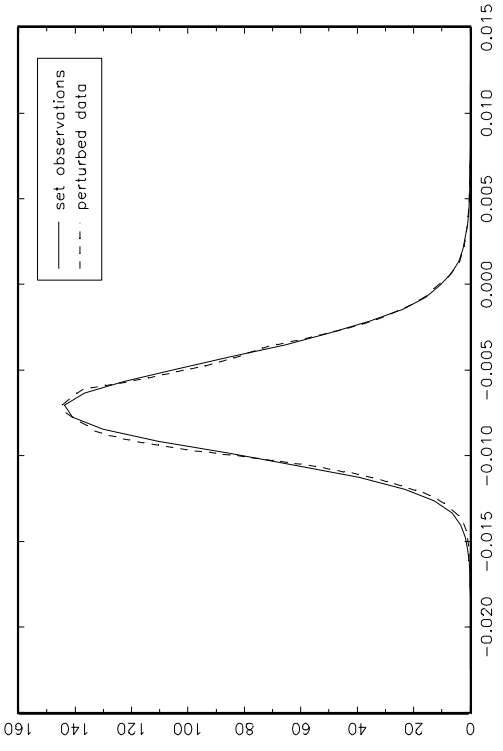

Figure 2: Posterior Density for mu

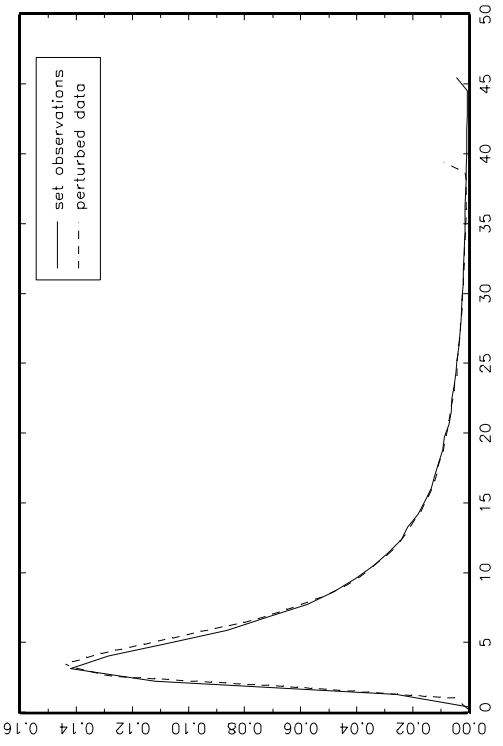

Figure 5: Posterior Density for nu

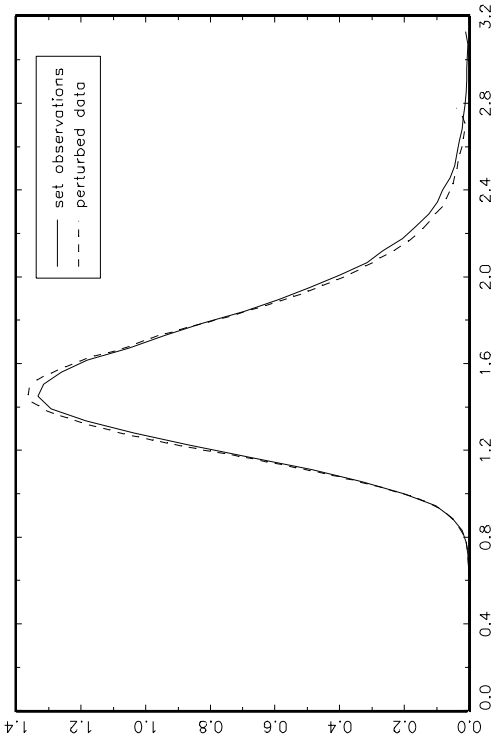

Figure 4: Posterior Density for gamma

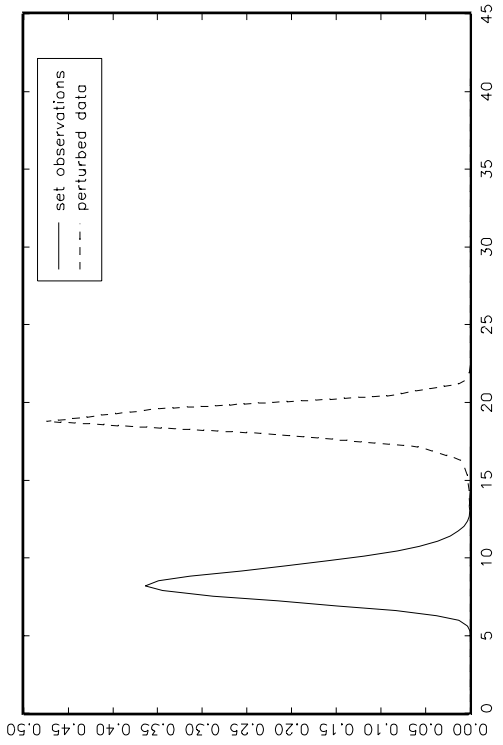

Figure 7: Posterior Density for ln(tau), nu=0.1

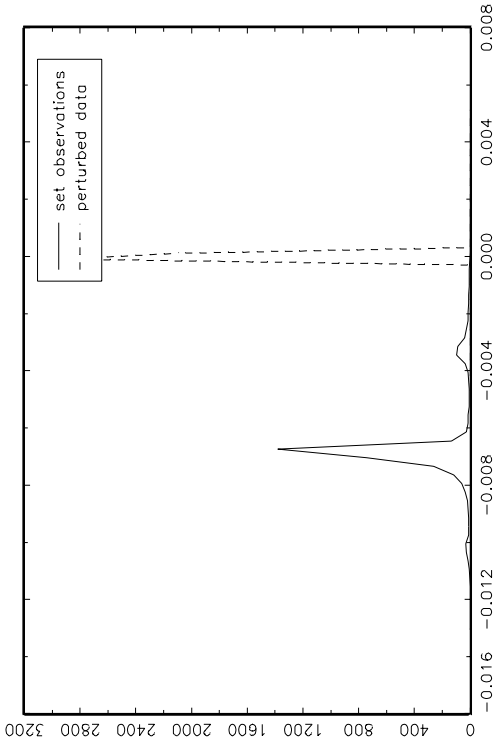

Figure 6: Posterior Density for mu, nu=0.1

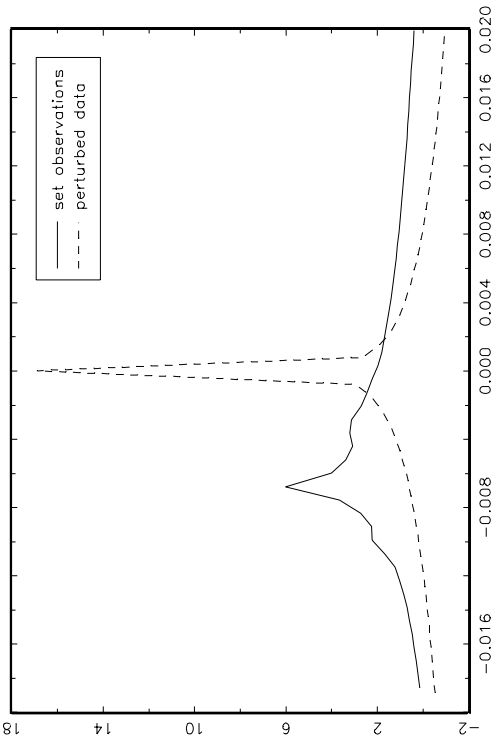

Figure 9: Log of Predictive Density, nu=0.1

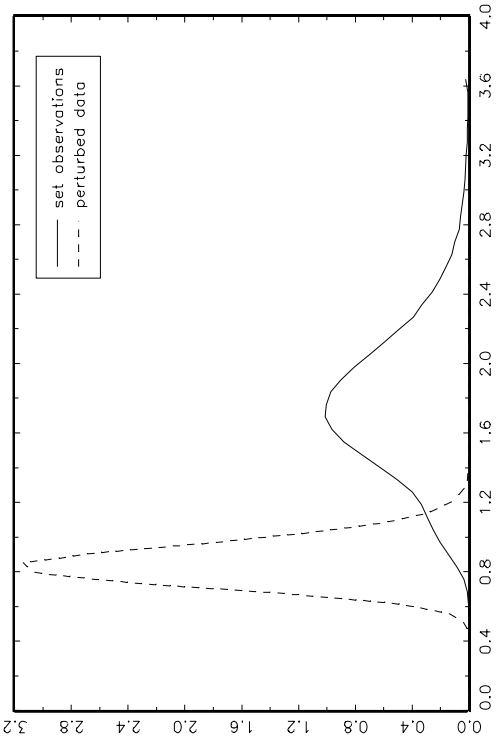

Figure 8: Posterior Density for gamma, nu=0.1